Appendix 7. Sample Allocation Methodology for Commodities and Services

Introduction

The primary objective of the Commodities and Services (C&S) sample design is to determine values for all sample design variables that minimize the sampling variance of 6-month price change for the C&S portion of the Consumer Price Index (CPI). The sample design variables are the number of entry level items (ELIs) to select in each item stratum and the number of outlets to select per Telephone Point-of-Purchase Survey (TPOPS) category-replicate panel in each Primary Sampling Unit (PSU). To that end, the variance of price change for the C&S portion of the CPI and the total annual cost of data collection and processing are modeled as functions of the design variables. These models allow the sample design problem to be expressed as one of minimizing the total variance of price change, subject to various cost and sample allocation constraints. Within this framework, nonlinear programming methods are used to solve the problem for optimal values of the sample design variables.

Certain simplifying assumptions are made to render the problem tractable and operationally more manageable. The number of PSUs , the number of replicate panels per PSU, and the classification of ELIs into item strata have been determined in previous work (Williams et al., 1993; Lane, 1996). Item strata are divided into 13 item groups for the design: 4 food at home groups (nonmeat staples; meat, poultry, and fish; fruits and vegetables; and other food at home and alcoholic and nonalcoholic beverages); food away from home; household furnishings and operations; fuels and utilities; apparel; transportation less motor fuel; motor fuel; medical care; education and communications; and recreation and other commodities and services. The 87 PSUs are divided into 15 groups according to size and number of replicate panels. (See table 7.) It is assumed that the same item and outlet sample sizes will apply to all PSUs within the same PSU group. This reduces the allocation problem to one of determining the number of ELI selections per replicate panel by PSU group and item group $\{K_{ij}, i = 1,...,15, j = 1,...,13\}$, i = PSU group, j = item group, and the number of outlet selections per TPOPS category per replicate by PSU group and item group $\{M_{ij}, i = 1,...,15, j = 1,...,13\}$. These are the design variables.

Let s^2_{Total} be the total price change variance for the C&S portion of the CPI, and let C_{Total} be the total annual cost of data collection. Then, the sample design problem can be expressed as one of minimizing s^2_{Total} subject to the following cost and sample allocation constraints:

Table 7. PSU Groups for C&S Design

PSU Group	Name	PSU Group	Name
1	New York City	10	Non-self-representing PSUs, Census Region 1
2	New York City suburbs		
3	Los Angeles City	11	Non-self-representing PSUs, Census Region 2
4	Los Angeles suburbs		
5	Chicago	12	Non-self-representing PSUs, Census Region 3
6	Philadelphia and San Francisco	13	Non-self-representing PSUs, Census Region 4
7	Detroit and Boston		
8	Other large self-representing PSUs	14	Smaller non-self-representing PSUs, Census Regions 1-4
9	Smaller self-representing PSUs	15	Anchorage, AK, and Honolulu, HI

$C_{Total} \leq$ Total data collection budget for C&S

$M_{ij} \geq 2$, i = 1,..., 15, j = 1,...,13

$K_{ij} \geq$ Number of item strata in PSU group i, item group j, i = 1,..., 15, j = 1,...,13

$K_{ij} \leq$ Maximum number of item hits in PSU group i, item group j, i = 1,..., 15, j = 1,...,13

Average number of item hits per stratum-index area in in PSU group i, item group ≥ 9, i = 1,..., 15, j = 1,...,13

A detailed description of these methods follows.

The Sampling Variance Function

Variance components models attempt to allocate parts of the total sampling variance to different sources of variation. For the C&S item-outlet sample, the following four sources of variation are modeled: PSU selection, item selection, outlet selection, and a residual component that includes other sources, such as sampling within the outlet.

The variance function for the C&S sample design is modeled for index areas. Each self-representing PSU is a single index area. Non-self-representing PSUs represent 7 index areas, with the sample for each area represented by 2 to 22 PSUs. As mentioned above, the variance model assumes that the total variance of price change for item group *j* within index area *k* can be expressed as a sum of four components:

$s^2_{j,k} = s^2_{psu,j,k} + s^2_{item,j,k} + s^2_{outlet,j,k} + s^2_{error,j,k}$, where

$s^2_{j,k}$ is the total variance of price change for item group *j* in index area *k*,

$s^2_{psu,j,k}$ is the component of variance due to sampling PSU's in non-self-representing areas,

$s^2_{item,j,k}$ is the component of variance due to sampling of ELIs within item strata,

$s^2_{outlet,j,k}$ is the component of variance due to sampling of outlets,

$s^2_{error,j,k}$ is a residual component of variance which includes the final stage of within-outlet item selection , called disaggregation

Similarly, it is assumed that the variance of price change of an individual sampled unit or quote has the same structure:

$$\sigma^2_{unit,j,k} = \sigma^2_{unit,psu,j,k} + \sigma^2_{unit,item,j,k} + \sigma^2_{unit,outlet,j,k} + \sigma^2_{unit,error,j,k},$$

where

$s^2_{unit,j,k}$ is the total variance of price change of an individual sampled unit or quote for item *j* in index area k,

$s^2_{unit,psu,j,k}$ is the component of unit variance due to sampling PSU's in non-self-representing areas,

$s^2_{unit,item,j,k}$ is the component of unit variance due to sampling of ELIs within item strata,

$s^2_{unit,outlet,j,k}$ is the component of unit variance due to sampling of outlets, and

$s^2_{unit,error,j,k}$ is the corresponding residual component of unit variance

Thus the projected sampling variance for a given index area k in PSU group i is:

$$s^2(PC_k) = \sum_{j=1}^{13} RI^2_{j,k} \left(\frac{s^2_{unit,item,j,k}}{f_1(M_{ij},K_j,N_k)} + \frac{s^2_{unit,outlet,j,k}}{f_2(M_{ij},K_j,N_k)} + \frac{s^2_{unit,error,j,k}}{f_3(M_{ij},K_j,N_k)} + \frac{s^2_{unit,psu,j,k}}{f_4(N_k)} \right)$$

where

$f_1(M_{ij},K_{ij},N_k) = (N_k H_k K_{ij})$

$f_2(M_{ij},K_{ij},N_k) = [(N_k H_k M'_{i,j} + N_k H_k M_{i,j} NPV_j)\, NRO_j]$

$f_3(M_{ij},K_{ij},N_k) = [(N_k H_k M_{ij} K_{ij})(NRQV_j)]$

$f_4(N_k) = N'_k =$ *the number of non self-representing PSU's in index area k*

and

N_k is the number of PSU's in index area k ,

N'_k is the number of non-self-representing PSU's in the index area,

H_k is the number of replicate panels per PSU in the index area

NRO_j is the outlet initiation response rate for major group *j*.

$NRQV_j$ is the quote level response rate for major group *j* for variance projection.

NPV_j is the weighted sum of nonPOPS categories in major group j, each category weighted by its probability of selection, for variance projection

M'_{ij} is the number of unique in-scope outlets selected per PSU-replicate, modeled as a quadratic function of the outlet sample size:
$M'_{ij} = (AV_{ij} M_{ij} + BV_{ij} M^2_{ij})$

And the sampling variance of price change for the U.S. City Average C&S index is

$$s^2_{TOTAL} = \sum_j \sum_k RI^2_{j,k} w^2_k s^2_{j,k}$$, where

$RI_{j,k}$ is the relative importance of item group *j,* in index area *k*, scaled to sum to 1.0 over all C&S item groups, and

w_k is the 1990 Census population weight of index area *k*

Relative importances of item groups are obtained from the most recent two years of the Consumer Expenditure Survey. They are the proportion of total expenditures in index area *k* that come from item group *j.*

The Cost Function

The costs of the C&S portion of the CPI which are modeled are the costs of initiation data collection and travel, and pricing data collection (personal visit and telephone) and travel. Each of these models are developed in terms of outlet- and quote-related costs and as functions of the design decision variables.

Initiation Costs

Outlet Related Initiation Costs

For PSU group *i* and major group *j*, outlet related costs for initiation are:

$$CI_O(M_{ij},K_{ij}) = 0.25 N_i \cdot H_i \cdot (CO_j + COT_j) \cdot (AC_{ij} M_{ij} + BC_{ij} M^2_{ij} + NPC_j M_{ij})$$

$CI_O(M_{ij},K_{ij})$ is the outlet-related initiation cost for major group *j* in PSU group *i*

N_i is the number of PSU's in group *i*,

H_i is the number of replicates per PSU in PSU group *i*,

CO_j is the compensation initiation cost per outlet for major group *j*,

NPC_j is the weighted sum of nonPOPS categories in major group j, each category weighted by its probability of selection

COT_j is the per diem and mileage cost per outlet for major group *j*

and $(AC_{ij}M_{ij} + BC_{ij}M_{ij}^2)$ is an overlap function used to predict the number of unique sample outlets, accounting for the overlap of elements in the outlet sample within and between major groups for a replicate panel. The number 0.25 accounts for the rotation or reinitiation of the outlet sample in one fourth of the sample TPOPS categories-PSU's each year.

Quote Related Initiation Costs

Quote related initiation costs are

$$CI_Q(M_{ij}, K_{ij}) = 0.25 N_i H_i \cdot WOD_j \cdot CQ_j \cdot M_{ij} \cdot K_{ij} \cdot NRO_j,$$

where

$CI_Q(M_{ij}, K_{ij})$ is the quote-related cost of initiation for major group *j* in PSU group *i*,

WOD_j is a seasonal items initiation factor for major group *j*,

CQ_j is the initiation cost per quote for major group *j*

Repricing Costs

The costs of ongoing price data collection and processing are also developed as both outlet and quote related costs.

Outlet Related Repricing Costs

For PSU group *i* and major group *j*, outlet related costs for ongoing pricing are:

$$CP_O(M_{ij}, K_{ij}) = MBO_{ij} \cdot N_i \cdot H_i \cdot NROj \cdot (AC_{ij}M_{ij} + BC_{ij}M_{ij}^2 + NPC_j M_{ij}) \cdot [(CPVO_j + CPO_j) \cdot (1 - RTO_j) + CTO_j \cdot RTO_j],$$ where

$CP_O(M_{ij}, K_{ij})$ is the total outlet-related cost for ongoing pricing for major group *j* in PSU group *i*

$CPVO_j$ is the compensation cost (time spent in travel) for a personal visit for pricing per outlet for major group *j*,

CPO_j is the travel cost (per diem and mileage) for a personal visit for pricing per outlet for major group *j*, = 3.33 for every *j*

RTO_j is the proportion of outlets priced by telephone for major group *j*,

CTO_j is the per outlet cost for telephone collection, = 3.43 for every *j* initially,

NPC_j is the weighted sum of NONPOPS categories in major group *j*, each category weighted by its probability of selection for cost projections.

MBO_{ij} is a factor to adjust for the monthly/ bimonthly mix of outlets by PSU and major group.

Quote Related Repricing Costs

Quote related costs for ongoing pricing are:

$$CP_Q(M_{ij}, K_{ij}) = MBQ_{ij} \cdot N_i \cdot H_i \cdot M_{ij} \cdot K_{ij} \cdot NRQC_j \cdot [CPVQ_j \cdot (1 - RTQ_j) + CTQ_j \cdot RTQ_j],$$ where

$CP_Q(M_{ij}, K_{ij})$ is the total quote-related cost for ongoing pricing,

MBQ_{ij} is a factor to adjust for the monthly/ bimonthly mix of quotes by PSU and major product group.

$CPVQ_j$ is the per quote cost (compensation not spent in travel) for a personal visit for pricing,

RTQ_j is the proportion of telephone collected quotes for major group *j*,

CTQ_j is the per quote cost for telephone collection for major group *j*, and

$NRQC_j$ is the quote level response rate for project ing costs for major group *j*.

Total Cost Function

The total cost function associated with data collection for C&S, summed over all item groups and PSU groups, is then given by

$$C_{Total} = \sum_{i,j} [CI_O(M_{ij}, K_{ij}) + CI_Q(M_{ij}, K_{ij}) + CP_O(M_{ij}, K_{ij}) + CP_Q(M_{ij}, K_{ij})]$$

And the sample design problem can be expressed as that of minimizing the total variance, s^2_{Total}, subject to the constraints

$C_{Total} \leq$ *Total expenditure limit*

$M_{ij} \geq 2$ *i=1,..., 15, j=1,...,13*

$K_{ij} \geq$ *Number of item strata in PSU group i, item group j, i=1,..., 15, j=1,...,13*

$K_{ij} \leq$ *Maximum number of item hits in PSU group i, item group j, i=1,..., 15, j=1,...,13*

Average number of item hits per stratum-index area in PSU group i, item group i, ≥ 9, i=1,..., 15, j=1,...,13

We note here that the last set of constraints are added to address concerns regarding small sample bias at the elementary index level by assuring a minimum average sample allocation of nine expected quotes total per index area – item stratum combination.

Model Coefficients

The parameters of the cost function are estimated using agency administrative records, dating from fiscal year 1996 forward, and a Time and Travel Study conducted by the Office of Field Operations (OFO). Distinctions between personal visit and telephone collection of data are made based upon information from OFO and from an analysis of C&S microdata conducted within the Prices Statistical Methods Division. Response rates for each item group derive from field initiation records and ongoing pricing experience.

Since outlet samples are selected independently for each TPOPS category, and outlets may be listed in the sample frames for more than one TPOPS category, an individual outlet may be selected more than once. For example, a grocery store could be selected for both bakery products and dairy products. Thus, the number of unique outlets realized by the sampling process is needed to project outlet-related costs. Quadratic regressions are used to predict the number of unique outlets realized in sample selection as a function of designated sample size. These are developed and reevaluated with each rotation by modeling the number of unique outlets obtained in simulations of sampling procedures for each PSU and item group as a function of designated sample sizes, using the most current sampling frames available for each item .

Components of price change variance are computed using restricted maximum likelihood estimation methods with C&S price microdata, the most recent estimates being based on price data collected in 1998-2000. Component estimates are developed for 6-month price change for the 13 item groups for each index area and month. Mean unit components of variance estimates are then computed by averaging the unit components of variance across months.

Problem Solution

Solutions are found using three methods: the SPLUS NUOPT code, which utilizes a trust region method, SAS PROC NLP, which uses a quasi-Newton algorithm, and a SAS PROC IML procedure which also employs a quasi-Newton algorithm. In practice, each method has yielded identical solution sets. For each item group, the number of item selections is bounded below by the number of strata in the item group and above by a ceiling of 140% of the item group's previous item sample allocation.

ELI selections are then distributed among item strata within each item group, with consideration given to differences in relative importance, production stratum-level price change variance estimates, and response rates among the item strata within each item group, as well as special problems identified by commodity analysts and field staff. Similarly, designated outlet sample sizes are distributed among the various TPOPS categories in item groups to manage variation in expected response rates and respondent burden.

In general, recent sample designs have shifted resources in many item groups from sampling many outlets to fewer outlets, with more item selections per outlet. This is due primarily to the large residual component of price change sampling variance estimated for most item groups, coupled with an increasing trend in the number of unique outlets realized in TPOPS sampling.